GLM-4.5-Air Benchmark Performance Across Reasoning Mathematics and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of GLM-4.5-Air on reasoning mathematics coding and language understanding tasks. 12 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: AutoMonitor-Bench: Evaluating the Reliability of LLM-Based Misbehavior Monitor. Research question: What are the benchmark performance scores of GLM-4.5-Air on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.0/10.

3 Results

10 papers retrieved. 12 claims extracted; 12 independently verified. Quality review score: 9.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
AutoMonitor-Bench is the first benchmark designed to systematically evaluate the reliability of LLM-based misbehavior mo	✓	0.41
AutoMonitor-Bench consists of 3,010 carefully annotated test samples spanning question answering, code generation, and r	✓	0.39
Monitors are evaluated using two complementary metrics: Miss Rate (MR) and False Alarm Rate (FAR).	✓	0.25
Miss Rate (MR) captures failures to detect misbehavior.	✓	0.18
False Alarm Rate (FAR) captures oversensitivity to benign behavior.	✓	0.19
12 proprietary and 10 open-source LLMs were evaluated using AutoMonitor-Bench.	✓	0.21
There is substantial variability in monitoring performance among the evaluated LLMs.	✓	0.15
There is a consistent trade-off between Miss Rate (MR) and False Alarm Rate (FAR), revealing an inherent safety-utility	✓	0.29
A large-scale training corpus of 153,581 samples was constructed to fine-tune Qwen3-4B-Instruction.	✓	0.26
Fine-tuning Qwen3-4B-Instruction on known, relatively easy-to-construct misbehavior datasets was investigated to improve	✓	0.28
The results highlight the challenges of reliable, scalable misbehavior monitoring.	✓	0.24
The results motivate future work on task-aware designing and training strategies for LLM-based monitors.	✓	0.31

References

- https://openalex.org/W7123371870
- https://doi.org/10.48550/arxiv.2404.04167
- https://doi.org/10.1561/2200000087